Predicting the UK General Election using Twitter and Data Mining

By: *Harry Coultas Blum*
Supervisor: *Dr Aris Perperoglou*

Table of Contents

Introduction

The purpose of this project was to investigate the possibilities of using Twitter to predict the United Kingdom's general election in May 2015. Twitter has become a powerful tool for researchers in a variety of fields. This project will demonstrate some of the techniques that are being used today to get a rough estimate of what people are thinking.

A general election in the UK is decided through a voting system called 'First Past the Post'. Essentially, areas in England, Scotland, Wales and Northern Ireland are split into 650 parts where each constituency contains roughly the same amount of citizens inside it. These constituencies each have a seat in the House of Commons. Candidates who win the seat of constituency become a Member of Parliament (MP), with the purpose of representing the citizens from that constituency. Each candidate often belongs to a larger group of candidates and supporters known as a political party. The party with the most seats (not majority) has "won" the election giving their cabinet, their party leader and set of MP's, the opportunity to pass bills (laws). As the party that wins holds a majority of the 650 seats, they have a more influence on what laws are passed. In a few circumstances, a 'hung parliament' is created. This is a situation in which a party with the largest amount of seats seeks to find an ally from a smaller party. The government of the last five years in Britain has been characterised by a Coalition of two political parties to form a government. This is an outcome that is a distinct possibility for the May 2015 election. *(Wikipedia)*

Twitter has become a platform of social discussion. With over one million tweets per hour, over 300 million tweets sent since Twitter was created with the average user spending approximately 170 minutes per month on Twitter, the wealth of data, for possible analysis, that Twitter has is incredible. It is used by many different types of institutions as a way of communicating and creating an interface with the public. Radio and television programs use it as a fast way of interacting with the public for song requests and competitions. Businesses use it in adverts to perform market research. People in the public eye use it to let fans and followers know what they are doing as well as responding to requests and queries. Because of this popularity, research on Twitter over the last few years has been extensive; increasing as social media became more popular.

The project involves the analysis of large amounts of information that otherwise by hand would be extremely difficult to fully analyse. Using small statements made by users on the social media platform Twitter combined with a analysis of whether the tweets were positive or negative, the project attempted to predict the result of the election in 2015. The term for this is Data Mining and is related to subject of Big Data. Big Data has become a huge area that industries are trying to utilise in order to make a more accurate and informed decisions about their business plans by looking at vast amounts of information to predict people's interests, behaviour or opinions.

People who have decided to follow users on Twitter will have tweets from these users shown on their home timeline. Tweets on a users personal home timeline are ordered by most recent and are the first thing a user when they sign in to their account. Many of the users on Twitter forward messages from other users to their 'followership', the Twitter term for this being 'retweet'. Tumasjan (2010) found that 19% of all tweets from their sample of 104,003 tweets were retweets.

The micro-blogging platform is not only a declaration of opinion to followers but seems to be a means of conversation. Honeycutt and Herring (2009) discovered that from a random sample of tweets, 31% use the @ sign and 91% of these were directed at a specific address. A small proportion of these were used to mention an addressee in a statement. This suggests that Twitter is a platform of discussion and conversation.

Smith (2011) has shown that 22% of adults were actively engaging with political campaigns through Twitter, Facebook and other social media platforms, leading up to the November 2010 US elections Clearly using Twitter, as a means of analysing public opinion over politics, could be very effective as the statistical population sample involved in Twitter is significantly bigger than the size sampled in polls.

Many researchers suggest that Twitter is a good means of predicting the election. Tumasjan's and Andranik's (2010) study of the elections in Germany was relatively new for its time. The main aim of the research was to see whether "micro blogging messages can actually inform us about the political landscape in the offline world". They found that the mere number of tweets reflects voter preferences, similar to the suggestion Véronis (2007) made about mentions of a party in the press being a better predicator of electoral success than the polls. Notably, the sentiment of the tweets was found to be similar to that of the current political sentiment in the media for this particular study.

Boutet, Hyoungshick, Eiko (2012) looked at the UK general election in 2010, attempting to go a step further and profile users based on their interaction with manually profiled users. They quantified the activity in several ways including the standard number of mentions and retweets related to a given topic but also some Bayesian probability to guess how likely it was that a user was affiliated to a party, given the amount of interaction that they have had. They assumed that that all users found were a member of one of the main 3 parties and took the highest probability of the 3 as the party the user affiliated with. The Bayesian-Volume produced 86% accuracy when compared to the users whose affiliation was already known. Interestingly, they also used a method called Support Vector Machine as proposed in Pennachhiotti and Popescu (2011). This used sentiment analysis amongst general user statistics, to calculate the average emotion over the users' tweets. This method produced only 62% accuracy suggesting that it isn't effective.

There are critics that think that using Twitter to analyse public opinion, in regards to politics, isn't ever likely to be an effective or reliable methodology. Avello (2012), has looked at others research that has attempted to estimate political opinion. He makes the point that not everybody is using Twitter. Therefore the demographic being used is limited compared to the polls, which have a much wider and more targeted demographic. Avello also states that just because a user is saying it on twitter does not mean that it's true. A lot of the data being collected on twitter is not going to be accurate and Avello suggests that it should be discarded. Avello makes several more suggestions that could increase the reliability of election prediction through Twitter. He states that sentiment analysis in politics should be thoroughly specialised for the topic before any predictions are made. Mainly because there is a lot of jargon, humour and sarcasm in politics that needs to be detected accurately by a sentiment analysing tool. Otherwise results may by completely inaccurate. Also he promotes the need for a definition of how the researcher plans to count votes so that other researchers can trust the methods being used. Overall the suggestions he made are completely appropriate however when doing a study on whether Twitter is a good resource to predict elections alone, it would be extremely difficult to follow all of these as an individual engaged in a small scale research project.

Novel Method

The main aim of this project was to estimate the 2015 May general election results as accurately as possible. Like some of the studies mentioned previously e.g. *Tumasjan's and Andranik's (2010)*, the comparison of this projects estimation against the more traditional polling method, would be the key measure of success. However it made sense to also look at each constituency independently to find out whether the results were correct. This is due to the type of voting system used in the UK, the percentage of votes a party gets does not always translate into seats. For example in the 2010 general election, the Liberal Democrats got 23% of the votes however they only received 8% of the seats (http://en.wikipedia.org/wiki/United_Kingdom_general_election,_2010).

Location

After reading research with similar aims, it was clear that no one had entirely taken into account the user's locations in their research. The location of a citizen in the United Kingdom is a key factor that could be studied in relation to the voting system used here. Twitter's Application Programming Interface allows developers to search for tweets in a specific area.

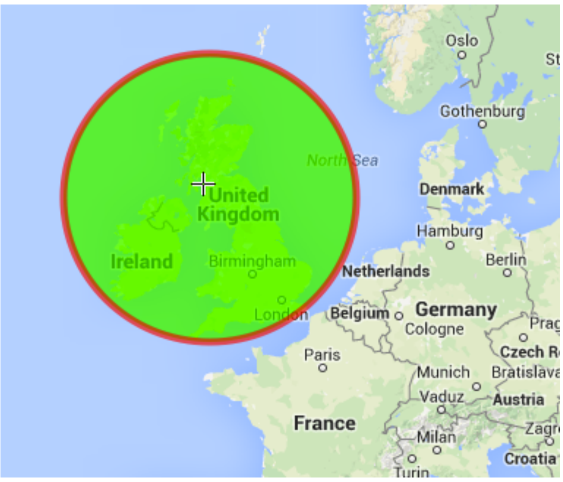


Figure 1 – Area used in the program

Before implementing this feature, searching for terms like '#Conservative' would have sent back a huge amount of tweets based in America and other locations. Although this information could be useful in gathering important insight into global opinion, which does have some value when trying to predict the UK elections, it is more likely to distort the opinion estimation of the United Kingdom's electorate. Also, searching for locations in the general vicinity of *Figure 1* ensured that tweets without location were not being included into the data set. This was important as without location, tweets give very little information about the person writing behind them. Compared to the range of demographic information produced by polls, this would have been an extremely limiting factor. Location is arguably the most important demographic element in regards to the general election. This is because seats in the House of Commons are divided up geographically across the country as explained earlier. Therefore forcing the miner (data collector) to find tweets with locations gave the results a better chance of competing against the poll predictions and more importantly the actual results.

The Twitter API

For this project, the main resource for collecting information was the Twitter API, which allows 180 queries every 15 minutes. In order to fully take advantage of this, a program written in Java was compiled onto a RasberryPi. *(You can find the source code for the program in the Appendix).*

This device was then left to make queries every 15 minutes based on keywords in a text file. Once the tweets had been collected, information about them was evaluated. The sentiment of the tweet was analysed based on chosen keywords associated with the main political parties. User profiles were created with the average sentiment towards parties mentioned for that user.

After downloading two text files from the RasberryPi, one containing the information about each user and the other, containing all the tweets collected, it was then possible to analyse the data in R.

User Profiling

The next important step was to profile users based on tweets that had already been obtained. Once an individual tweets once regarding any of the parties, it was possible to try and work out whether they had mentioned this party with a positive or a negative sentiment. Using the Stanford Universities core Natural Language Processing library for Java, it was possible to do this.

The library essentially works by tokenizing the text given to it, tagging the words with a Part of Speech Tagger and breaking up each part of the sentence into a tree. The process is then completed when these trees are compared to the models in the tree bank provided by the library in which the end result of a sentiment is given. Each sentence model has been reviewed by 3 judges in order to achieve high accuracy (Socher, 2013). Therefore the tool, given a string of text, provides users with a powerful estimation of sentiment. Statistics suggest that this model is in fact correct 85% of the time.

Figures 2 and 3 show examples of the libraries capability and the results that it can produce. '++' suggests a very positive sentiment while '--' is very negative. For the project, these values are 5 and 1 respectively.
After the tweets were downloaded the sentiment of the tweet was analysed using the Stanford Natural Language Processing library in Java.

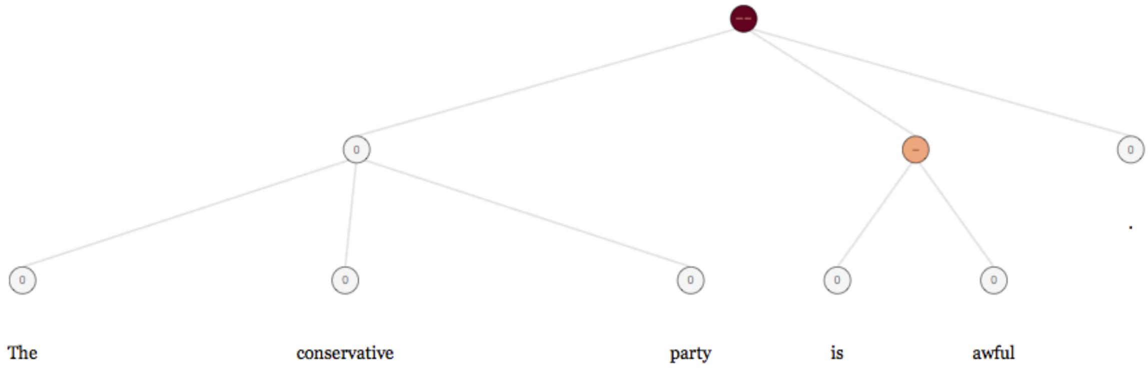


Figure 2

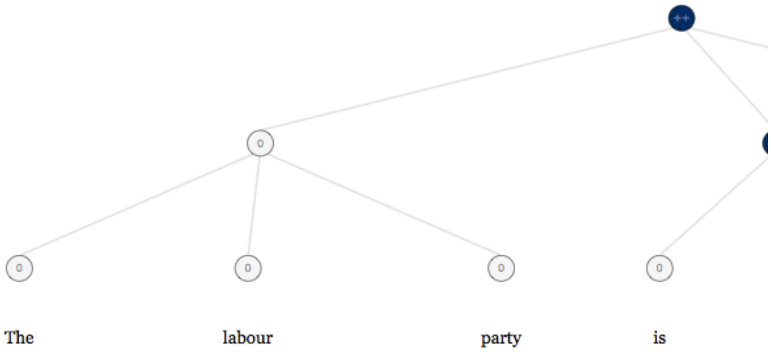


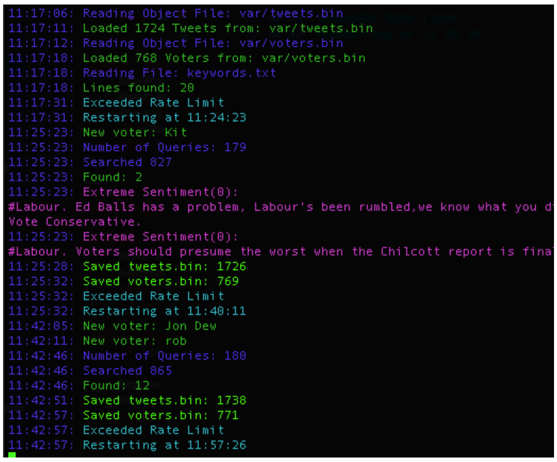
Figure 4

Figure 3

The statement "The labour party is amazing" is analysed as extremely positive. If this were to be an actual tweet, the analysis of this would give a lot of insight into how that particular user who said it was thinking about voting in the general election. Combined with their location, this could yield an extremely interesting set of predictions.

Implementation

Implementing the tweet miner in Java imposed a number of organisational challenges. The most important issue was to set up a way of allowing the program to consistently hit Twitter's API in order to get any new tweets.



```
11:17:06: Reading Object File: var/tweets.bin
11:17:11: Loaded 1724 Tweets from: var/tweets.bin
11:17:12: Reading Object File: var/voters.bin
11:17:18: Loaded 768 Voters from: var/voters.bin
11:17:18: Reading File: keywords.txt
11:17:18: Lines found: 20
11:17:31: Exceeded Rate Limit
11:17:31: Restarting at 11:24:23
11:25:23: New voter: Kit
11:25:23: Number of Queries: 179
11:25:23: Searched 827
11:25:23: Found: 2
11:25:23: Extreme Sentiment(0):
#Labour. Ed Balls has a problem, Labour's been rumbled,we know what you d
Vote Conservative.
11:25:23: Extreme Sentiment(0):
#Labour. Voters should presume the worst when the Chilcott report is fina
11:25:28: Saved tweets.bin: 1726
11:25:32: Saved voters.bin: 769
11:25:32: Exceeded Rate Limit
11:25:32: Restarting at 11:40:11
11:42:05: New voter: Jon Dew
11:42:11: New voter: rob
11:42:46: Number of Queries: 180
11:42:46: Searched 865
11:42:46: Found: 12
11:42:51: Saved tweets.bin: 1738
11:42:57: Saved voters.bin: 771
11:42:57: Exceeded Rate Limit
11:42:57: Restarting at 11:57:26
```

Figure 4 gives a general idea of what the program looks like when it is running. It saves tweets and voters into a file as it goes along and reports any interesting finds like Extreme sentiments as well as new voters. It was possible to transfer the results of the searching through the SCP protocol and then analyse this information in R.

If a tweet in the query was found that was not already in the dataset it was added and its sentiment was analysed. If the tweet was by a user that hadn't already been seen, a new "Voter" element was created. If the user was already in the set of voters already seen, the sentiment of the tweet that was found would be calculated and depending on the search query, the voter's average sentiment of tweets related to that query, would also be calculated. For example one tweet obtained from David Stuart Cole, is in the table below. None of the parties mentioned have completely positive sentiment (>3). This was an issue throughout the project as the majority of tweets were evaluated to have negative sentiment. However we can guess that David is going to vote for Labour as that is the closest to being neutral (3). As the number of tweets is quite high, the result 2.5 suggests that there was a tweet with a sentiment of 2 or higher.

Name	**Screen Name**	**ID**	**Conservative**	**Labour**
David Stuart cole	Coldavsc66	2345690571	2	2.5
Green	**Lib-Dem**	**UKIP**	**Location**	**Number of tweets**
2	2	2.4	Tupton	29

One issue that needed to be prevented was the possibility of a user never making a tweet about the other parties. To resolve this problem if a user had not mentioned a party in and of their tweets, the value of the sentiment was set as 0 to visualise this.

As the majority of the time only one tweet for each user was being found, an additional step was implemented to the search. This was to ensure that as much information was found out about the users picked up by the tweet miner as possible.

This was performed as follows. While new users were being found, they were also added to a list. After hitting the first API limit, the miner searched through as many of the tweets of those new users as possible. If any of the tweets found matched the '#' or '@' tags highlighted in the keywords file, while also being tweeted in Great Britain, the tweet would be added and analysed. If a low amount of new users were found, the miner would try and look through users already gathered to see if they had any more relevant tweets on their timeline. Once a user's timeline had been completely scanned, the miner would never scan them again.

This process allowed the miner to catch up on previously skipped users so the most relevant data could be collected on them. The main idea behind this process was that users who have at least one tweet about relevant keywords, may have made reference in other tweets .

Starting from the 1st of March and finishing on 1st of April, the miner continued to profile users searching for tweets that matched the following overall criteria:

1. The tweet had a location

2. The location of the tweet was in an area where voting would occur

3. The tweet had keywords in it that matched the list of relevant keywords.

4. The tweet was by a user who had already made at least one tweet relevant to the keywords.

Tweet Example Table

Name	Query	Date	Location	Sentiment	Text
craig leishman	Conservatives	2014-12-24 20:55:08	East Midlands	5	*Happy Christmas to @Conservatives everywhere - if your politics are different now is a great time to jump on board with the winning team*
I mac	Conservatives	2014-11-10 08:04:35	Swansea	1	*Roads a mess for year's now 6 months from election @David_Cameron spends billions, do they think we're daft #torylies*
Ursula Gardner	UKIP	2015-03-16 15:48:02	Bristol	2	*Lesson from today's lecture: it's actually impossible to consider #ukip anything but racist*
Fabian Breckels	Labour	2015-03-26 22:25:30	Bristol	4	*@bevclack Dave knows he'll be ripped to shreads by @Ed_Miliband*

The main analysis of the tweets was done by looking at the sentiments of users tweets that had mentions of the five main political parties and their leaders:

The Conservative Party – David Cameron
The Labour Party – Ed Miliband
The Liberal Democrats – Nick Clegg
UK Independence Party – Nigel Farage
The Green Party - Natalie Bennet

Once the tweet-miner had collected and processed the tweets and the users tweeting them, we were left with information in the format seen in the "Tweet Example Table". This table shows how information was collected for 4 people. In the table the sentiment column tells us what the Stanford NLP sentiment analysis result produced (+1). For example the first tweet in the table was evaluated as 5 meaning that it was extremely positive. The second tweet in the table was evaluated as "1" meaning that it was extremely negative. "2" represents negative and "4" positive while "3" represents neutral. The reason that we added one to the results was so that we could plot the results on a star graph. Examples of these star graphs are in the results section (Results).

The next table I have included, titled *"Voters Example Table"*, illustrates nine examples of users collected. The columns for the political parties represent the average sentiment of tweets that have been directed towards that party. For example if we look at the first item in the table, a user named Trevor Forrester has tweeted 457 times, some of those tweets being about Labour with those tweets having an average sentiment of 2. Columns containing "0" mean that the user has not made any tweets about that specific party and therefore has no average sentiment. The "Name" and "Screen Name" columns represent a users name and display name respectively. Interestingly, most of the screen names of collected users could have been people's actual full names. Of course the Location column represents the number of tweets that were collected for that individual user.

An issue that arose during the project was that none of the average sentiments towards a party for a given user was positive. Guessing what party a user with only negative sentiments, was going to vote for is difficult as it doesn't tell us who they are in sympathy with. One option to diminish this issue was to take the party with the highest mean and closest to neutral. However this on its own didn't seem like the best solution to the problem.

In the end a weighted average was created as it was the best option. As the sentiment analysis rarely assessed tweets to be extremely negative or positive, it made sense to make the effect of them on the users overall sentiment higher. For example, if a tweet were analysed to have a sentiment of 5, it would be boosted up to 5.6 so that its affect on the overall average would be higher. By doing this, extremely negative tweets would cause the program to assess the users to be against that party even if they had a large amount of positive (sentiment = 5) tweets.

Voters Example Table

Name	*Screen Name*	*Conservative*	*Labour*	*Green*	*Lib Dem*	*UKIP*	*Location*	*Number of Tweets*
Trevor Forrester	Trev_Forrester	2.1458	2	2.5	2.25	2.1899	West Midlands	463
Ronnie	ClarkAndSon Ltd	2.0625	2.1905	0	2	2.1455	Wigan	305
Malcolm Carter UKIP	MalcCarter	2.75	2.1657	0	0	2.2411	West Midlands	285
craig leishman	CLeishman1969	2.3273	2.2588	0	2	2.3684	Bottesford	241
The Man in Black !	Roger_Sussex	2.1124	2	2	2.087	0	Treorchy	227
SeÃ¡n Robertson	SeanLXIV	1.8774	2.5714	0	2.1379	2.2222	Liverpool	209
brian hooper	8589brian	2.3067	2.4615	0	2.35	2.6667	Cheltenham	186
AndrewCB	ACBLive	2.1111	2.4444	0	2	2.162	Droitwich	180
Jeffrey Muir	muir_jeffrey	2.5	2.2857	0	0	0	Perth	176

Prelect

Predicting the General Election

http://himion0.github.io

The results section can be found on the website found by clicking the link above. This includes information about the sentiment analysis and general and location based predictions derived from the tweets. A printed copy can be found on the next page however this presentation is not the optimal way of looking at the results section.

Data Sets

Over the course of the study there were roughly 15,000 tweets collected. These tweets formed the first data set and helped us to make a general prediction of the general election.

One of the main aims of the project was to classify users. A data set was created containing the users who had made the tweets that we had already collected. These users therefore each had a set of tweets. The number of user collected after the month of research was around 3000.

As only tweets with locations were collected, it was possible to guess where this user was located. For example a user with 35 tweets overall, 30 of them being tweeted in Leeds, would be assumed to be located in Leeds. This algorithm was of course based on the mode i.e.

$$l_i = i^{th}\ location$$
$$where\ i = 1,2,\ldots,n$$

n is the number of unique locations that user tweeted from.

Then for a user U_j the function to find the location of this user would be:

$$Loc(U_j) = \max\,(l_1, l_2, \ldots, l_n)$$

This allowed us to group users based on where they were from and guess what the result of a specific constituency would be.

A similar algorithm was used to classify the users political opinions. Each tweet was given a sentiment derived from the Stanford NLP tool. These sentiments where then grouped where then grouped based on whether they were about the Conservatives, Labour party, Liberal Democrats, Green party or UKIP.

The average sentiment for each of these groups of tweets was calculated e.g. for the Green Party:

$$Green(U_i) = \frac{\sum_{j=1}^{n} sentiment(greentweet_j)}{n}$$

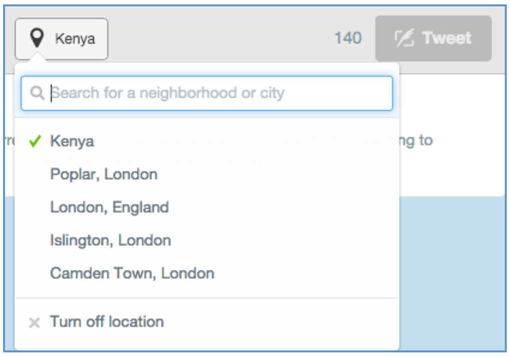
The prediction of what party a user was going to vote for, was the group with the highest average sentiment.

$$\max[Conservative(U_i), Green(U_i), Labour(U_i), UKIP(U_i), LibDem(U_i)]$$

Adding these results to the user data set, we were able to make a prediction of what a user was going to vote for and therefore what a specific locations result might be.

Drawbacks and Improvements

Throughout the project, I noticed several things that could be improved and make the algorithm for finding and analysing tweets more effective. I also realised areas in which the research process I had devised, could be ruined by a third party. One of these was that on Twitter it is possible to falsify location of tweets.



The picture above, highlights how it is possible to set your location as if you were from anywhere. However, it was more than likely that most of the users that I collected tweets from probably weren't doing this. The reason being that no one knows about the project and would not be seeking to undermine it. If the method were to become more mainstream this would be an issue. One possibility of getting round this would be to create a user profiling algorithm that tries to search for a proper address of a user. This in itself, would be a very difficult task, as it might require the creation of a web crawler and perhaps implementation of several additional Application Programming Interfaces.

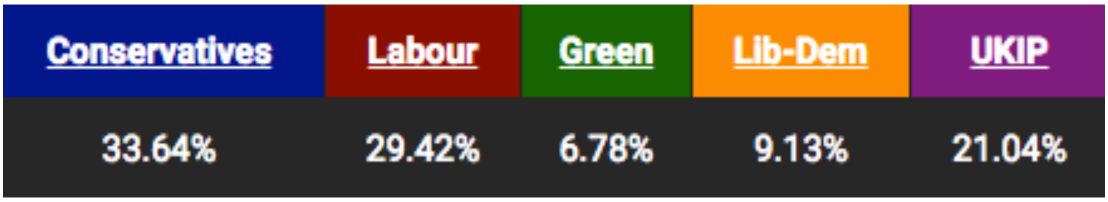
As mentioned before, Stanford's NLP's tool for sentiment analysis is very powerful. However it hasn't been designed to take into account the political satire involved in political conversation, one of the most difficult areas in language to understand. As we have seen it has made errors in the process of estimating sentiment. The subject of politics is linguistically complex and nuanced. Therefore as a result of these issues, it would be wise to focus on creating a tool specifically designed for the process of calculating sentiment of politically fuelled comments like the ones that have been collected in this project.

The process of calculating sentiment included the entire tweet text. This was seemingly effective as Twitter only allows for 140 characters in each tweet so the majority of tweets were single statements. Some however, included 2 or 3 statements and mentioned different parties. If the text was chunked, which involves breaking the statements up, and then subsequently analysed, we might have learnt more about users opinions.

An example of this can be seen in the following tweet where both the Labour and Conservative leaders are mentioned. If the tweets were chunked properly the calculation of a false sentiment may have been avoided.

"Looking forward to the leaders debate @David_Cameron vs @Ed_Miliband. Not giving my position away but...............COME ON YOU BLUES!!!!!!"
- casper2136

Although more than 15000 tweets and over 3000 users were found towards the end of the data collection period, it may have been more effective to widen the search keywords in order to increase the size of the data. This could have induced more grounded forecasts. An interesting way of doing this might be to get the program to look at keywords that users were including in their tweets and see if those keywords were popular in other tweets. Including these keywords in the set of queries would increase the scope of users to be profiled and increased prediction accuracy. For example, a tweet including #David_Cameron may also contain #gethimout. This hash tag could be key in finding whether users would like David Cameron in number 10 for another 5 years. Another issue, caused by the low number of users collected was that location based predictions weren't as accurate as they could be. Some locations for example, only had a few



Conservatives	Labour	Green	Lib-Dem	UKIP
33.64%	29.42%	6.78%	9.13%	21.04%

users in them. This made predicting the result of this area unreliable. In the results section you can see this problem by clicking on the "More Results" button.

In order to make the voter profiles more reliable, previous tweets from a users timeline were brought in regardless of date. The reason for this would be that getting more tweets increases the accuracy of user profiling and gives more insight into voter opinion. A potential drawback however is that this method may have injected the data with tweets too far back in the past that so irrelevant as to distort the overall picture of the data. On the other hand, traditional polls often include data as old as five years (Wikipedia) suggesting old tweets are still relevant and implementing a maximum age to tweets could easily be implemented to avoid the issue of tweet age becoming a serious impediment to the data.

Finally, the method of classification that was used to guess which party users aligned with, was not as sophisticated as it could have been. As described in the "Data Sets" section, the method used the mode of the mean sentiments for tweets from each of the parties. An improvement to this might be to implement unsupervised learning. This involves determining the hidden structure of some un-labeled data set (Wikipedia). Using this we could try and better estimate users political affiliations to the main parties.

Conclusion

After roughly a month of data collection, the project produced the prediction that the Conservative's would be the most popular out of the 5 main parties. However this achievement would be a hung parliament as the Labour Party would follow behind with UKIP a close third. The estimation of what parties would join together to form this Hung parliament is therefore the real question that has been generated from the project corroborating with what more traditional polling methods were predicting.

Throughout the project I have learnt that there is huge room for improvement of the program as a sophisticated and credible tool of analysis. The results that were produced weren't as inaccurate or different from the pre election opinion polls. This differed from natural scepticism when starting the project that the tool wouldn't produce these kinds of results. When there was a sufficient amount of tweets to analyse for each user, the predictions were fairly similar to polls, followed previous election patterns and the popular opinion of correspondents reporting in the main stream media.

Social media platforms like Twitter are becoming ever more popular forms of conversation, expression and information for people. This only increases the richness of quality data to be collected and potential voters to be profiled during the run up to an election or even referendum on whether the UK should leave the EEC.

There are many ways in which the project could be improved. Some of these include running the program over a longer period, with a larger number of keywords and a politically specialised sentiment analysis tool. If the project were to be continued with these improvements implemented, it could become an extremely effective tool during the run up to elections. Perhaps data mining and processing social media platforms in this way could eventually replace the traditional method of polling as a source of unbiased approximation. By this I mean that it isn’t vulnerable to human error or interference in the process of gathering data.